PRACE GEOGRAFICZNE, zeszyt 141

Instytut Geografii i Gospodarki Przestrzennej UJ
Kraków 2015, 7–24
doi: 10.4467/20833113PG.15.006.4058

TWITTER JAKO ŹRÓDŁO INFORMACJI GEOGRAFICZNEJ

Łukasz Beluch

The Twitter as a source of geographic information

Abstract: In recent years, social media has gained significant popularity. The popularization of portable devices has produced a significant number of posts on social media along with a geolocation. Twitter is especially interesting. Thanks to an open Application Programming Interface (API), it ensures free access to published content. Data obtained this way gives new possibilities for social and geographic research. The research literature emphasizes the usefulness of Twitter for the purpose of the monitoring of extreme phenomena. This paper presents the main trends in research on Twitter tweets and technical solutions enabling data downloading.

Keywords: social media, Twitter, geolocation

Zarys treści: Media społecznościowe w ostatnim czasie zyskały znaczną popularność. Popularyzacja urządzeń przenośnych sprawiła, że znaczna część wpisów na portalach społecznościowych posiada geolokalizację. Szczególnie wart zainteresowania jest serwis Twitter, który umożliwia, dzięki otwartemu *Application Programming Interface* (API), dostęp do publikowanych treści bez żadnych opłat. Tak pozyskane dane otwierają nowe możliwości badań społeczno-geograficznych. W literaturze światowej podkreśla się na przykład ich dużą przydatność na potrzeby monitoringu zjawisk ekstremalnych. W tym opracowaniu przedstawiono główne nurty badań dotyczących wykorzystania *tweetów*, a także przybliżono rozwiązania umożliwiające pobieranie danych.

Słowa kluczowe: media społecznościowe, Twitter, geolokalizacja

Wprowadzenie

Koniec pierwszego dziesięciolecia XXI wieku to okres intensywnego rozwoju internetu. Szczególnie wykorzystanie technologii Web 2.0 przyczyniło się do powstania licznych, interaktywnych portali społecznościowych, z dynamicznie budowaną treścią (O'Reilly, Battelle 2009). Część z nich nie przetrwała próby czasu i upadła (np. Grono.net; Blip.pl), bądź też stale zmniejsza się liczba ich aktywnych użytkowników (Myspace.com). Inne, takie jak Facebook lub Twitter, wręcz przeciwnie, ciągle pozyskują nowych członków, którzy codziennie na całym świecie publikują miliony wpisów.

W 2007 roku firma Apple wypuściła na rynek pierwszego smartfona iPhone 1, tworząc zapotrzebowanie na nowego rodzaju usługi w postaci dostępu mobilnego do portali internetowych i co się z tym wiąże, potrzebę tworzenia aplikacji mobilnych opartych na udostępnionym *Application Programming Interface* (API, Kaplan 2012). Rozpowszechnienie urządzeń przenośnych, które posiadają stały dostęp do Internetu przeniosło media społecznościowe z rzeczywistości wirtualnej do świata realnego (Sui, Goodchild 2011). Obecnie coraz więcej wpisów publikowanych jest właśnie za pomocą smartfonów lub tabletów, w różnych sytuacjach życia codziennego. Urządzenia te wyposażone są w odbiorniki nawigacji satelitarnej i wykorzystują metody triangulacyjne do określenia położenia względem nadajników GSM i WiFi, co pozwala na dodawanie lokalizacji do treści wpisu, najczęściej w postaci współrzędnych geograficznych, ale również przez określenie miasta, rejonu bądź państwa. Powstają tzw. geotagi, czyli metadane zawierające współrzędne, na podstawie których możliwe jest późniejsze lokalizowanie wpisów w przestrzeni geograficznej (Crampton i in. 2013).

Dynamiczny rozwój urządzeń przenośnych, głównie smartfonów i tabletów, powoduje, że danych posiadających geolokalizację będzie szybko przybywać (Hawelka i in. 2013; Alexander 2014). Obecnie twórcy aplikacji mobilnych, zwłaszcza mobilnych wersji portali społecznościowych, domyślnie dodają lokalizację do treści pisanych przez użytkownika. Także w przypadku fotografii współczesne aparaty cyfrowe są coraz częściej wyposażone w odbiornik nawigacji satelitarnej, i informacja o tym, gdzie dane zdjęcie zostało wykonane, jest zapisywana w metadanych zdjęcia (Hyvärinen, Saltikoff 2010). W przyszłości stanie się możliwe połączenie różnych środków przekazu w jeden system, co zbliży nas do koncepcji zaproponowanej przez Goodchilda (2007) *human-as-sensor*.

W badaniach społeczno-geograficznych pozyskanie wiarygodnych, ankietowych danych badawczych na odpowiedniej próbie respondentów wymaga przeprowadzenia czasochłonnych i kosztownych ankiet, których wynik nie zawsze jest wiarygodny, ponieważ respondenci mogą celowo lub podświadomie udzielać tendencyjnych odpowiedzi, na przykład wstydząc się przyznać do pewnych swoich zachowań, dlatego Goodchild (2007) wskazuje na konieczność poszukiwania nowych tech-

nologii do pozyskiwania i analizy informacji (*data mining*), w tym użycie mediów społecznościowych.

Rewolucją w badaniach społecznych określaną jako „czwarty paradygmat" jest pojawienie się tzw. *big data*, czyli dużych ilości danych tworzonych przez społeczeństwo w stosunkowo krótkim czasie i nowych możliwości ich analizy (Hey i in. 2009). Nie są to tylko zmiany ilościowe, ale także zmiany jakościowe, pozwalające, dzięki współczesnej technologii, na szybkie porównawcze (krzyżowe) przeanalizowanie dużych zbiorów danych z różnych obszarów i uzyskanie informacji nieosiągalnej wcześniej stosowanymi metodami (Kitchin 2013; Shelton i in. 2014). Według Andersona (2008) klasyczna metoda naukowa polegająca na stawianiu hipotez badawczych dobiega ku końcowi, przyszłość natomiast ma metoda, w której zbiera się dużą ilość różnych danych, używa wystarczającej mocy obliczeniowej komputerów i szuka wzajemnych zależności pomiędzy na pozór zupełnie różnymi zbiorami informacji. Analiza wpisów z portali społecznościowych może objawić prawdziwe zwyczaje i nawyki ludzi, ponieważ nie są oni świadomi udziału w badaniach. Dodatkowo, analizując wpisy mające geolokalizację, można tak prowadzone badania osadzić w przestrzeni geograficznej. Mamy więc do czynienia z tzw. informacją AGI – *Ambient Geographic Information* (Crooks i in. 2013), co można przetłumaczyć jako nieświadomą informację, czyli informację geograficzną tworzoną przez ludzi, którzy nie są świadomi możliwości jej wykorzystania. AGI należy odróżnić od wcześniej opisanych w literaturze danych dostarczanych świadomie przez użytkowników internetu, czyli *Volunteered Geographic Information* (VGI, Goodchild 2007), których klasycznym przykładem jest OpenStreetMap, gdzie dane pochodzą od wolontariuszy zaangażowanych w wytwarzanie informacji.

W Polsce największą popularność wśród portali społecznościowych zdobył Facebook (Megapanel PBI/Genius 2014), który jednak oferuje dużo mniejsze możliwości pobierania danych w porównaniu z Twitterem, stąd zainteresowanie autora tym serwisem. Twitter jest portalem społecznościowym, tzw. mikroblogiem założonym w Stanach Zjednoczonych Ameryki w 2006 roku przez Jacka Dorseya (2006). Twitter opiera się na 140 znakowych wiadomościach tekstowych (*tweetach*) z założenia podobnych do SMS-ów (140 znaków na treść i 20 na podpis autora). Ze względu na to ograniczenie treść *tweetów* jest pisana specyficznym językiem zawierającym skróty i akronimy, również używane adresy URL muszą zostać skrócone. Zarówno w USA, jak i w Europie Zachodniej Twitter cieszy się znaczną popularnością. W Polsce w marcu 2014 roku użytkowników Twittera było prawie 3 mln i ich liczba wzrasta (Megapanel PBI/Genius 2014). Twitter jest obecnie narzędziem przekazu informacji stosowanym przez znane osoby, między innymi przez polityków. Przełomowym momentem dla serwisu, podkreślającym jego pozycję i wpływ na problemy o zasięgu ponadregionalnym był rok 2009, kiedy Twitter został użyty jako środek komunikacji protestującej opozycji podczas wyborów w Iranie (Newsweek 2009), a następnie podczas arabskiej wiosny w 2011 roku (Christensen 2011).

Celem artykułu jest przedstawienie obecnego stanu wiedzy dotyczącego możliwości użycia Twittera jako źródła informacji geograficznej, przybliżenie polskiemu czytelnikowi światowej literatury z tego tematu, a także opis technologii umożliwiającej pobieranie *tweetów* przez API Twittera.

Twitter w badaniach geograficznych na świecie

Pomimo tego, że Twitter istnieje dopiero niecałe dziewięć lat, w literaturze światowej zostało już opublikowanych wiele badań dotyczących użycia Twittera jako źródła informacji geograficznej. Zook i in. (2010) podkreślają pozytywną rolę mediów społecznościowych, w tym Twittera przy kartowaniu skutków trzęsienia ziemi na Haiti. Twitter posłużył za platformę wymiany informacji pomiędzy wolontariuszami zaangażowanymi w akcję ratunkową. Również Cheong i Cheong (2011) zauważają zalety użycia Twittera w badaniach społeczno-geograficznych, takie jak możliwość pozyskania dużej liczby informacji od ludzi dotkniętych kryzysem w krótkim odstępie czasu. Wykorzystali oni Twittera do badania ludzkich reakcji związanych z powodziami, jakie zdarzyły się w Australii w latach 2010–2011. W zarządzaniu kryzysowym informacje z mediów społecznościowych, w tym z Twittera, mają jedynie wspomagać tradycyjne źródła informacji kryzysowej, a nie je zastępować. Podejmuje się próby integracji obu źródeł informacji. Przykładem mogą być opracowania Schade i in. (2011; 2012), którzy rozważali integrację wpisów z Twittera z tradycyjnymi źródłami i metodami do monitoringu środowiska podczas powodzi w Wielkiej Brytanii oraz podczas pożarów lasów w Europie. Autorzy ci powołują się na wcześniejszą pracę de Longueville'a i in. (2010), w której została opracowana koncepcja *Digital Earth Nervous System* (DENS). Głównym założeniem tej koncepcji jest porównanie systemów monitoringu środowiska i wczesnego ostrzegania przed zagrożeniami z ludzkim systemem nerwowym. Podobnie jak w ludzkim ciele, na które musi zadziałać bodziec rejestrowany przez receptor, a następnie przekształcany w odpowiednie uczucie, po czym w mózgu człowieka podejmowana jest odpowiednia reakcja, tak i dane z mediów społecznościowych, a także ze źródeł teledetekcyjnych oraz tradycyjnych pomiarów, mogą być wykorzystane tak samo. W każdym z przypadków musi wystąpić jakiś bodziec, aby sensor go zauważył. W przypadku mediów społecznościowych sensorami są użytkownicy Twittera, którzy rejestrują niecodzienne zjawisko, na przykład pożar lasu czy powódź. Aby wyłapać takie wpisy, potrzebne jest przetworzenie tej informacji, podobnie jak mózg człowieka analizuje docierające do niego bodźce i rozróżnia te istotne od mniej ważnych. W przypadku danych z Twittera będzie to aplikacja komputerowa wyłapująca po słowach kluczowych (*#hashtagach*) bądź w bardziej zaawansowanej postaci analizująca semantycznie wpisy zawierające treści powiązane z określonym tematem. Wpisy te następnie są nanoszone na mapę z odpowiednią flagą (przykładowo pożar

lasu) i porównywane z danymi teledetekcyjnymi wykrywającymi tzw. *hotspoty*. Jeżeli występuje duża korelacja pomiędzy miejscami pożarów wykrytymi różnymi metodami, to znacząco wzrasta prawdopodobieństwo, że nie jest to fałszywy alarm i należy wysłać odpowiednie służby (Shade i in. 2012). Badania Crooksa i in. (2013) dotyczące reakcji ludzi na trzęsienie ziemi na wschodnim wybrzeżu USA wykazały, że Twitter może być też narzędziem do ostrzegania przed tym zjawiskiem, oraz potwierdziły, że stanowi on cenne źródło informacji przy ocenie skutków trzęsienia ziemi oraz miejscach najbardziej zniszczonych. Crooks i in. (2013) udowodnili, że teoretycznie informacja o trzęsieniu ziemi może dotrzeć do osób znajdujących się w jego strefie wpływu szybciej, niż dotrze tam fala sejsmiczna. Wystarczy, że osoba w epicentrum trzęsienia opublikuje taki wpis w chwili wstrząsu, co przy prostej formie wiadomości na Twittrze i dostępności aplikacji na urządzenia mobilne jest wielce prawdopodobne. Poza tym, możliwe do wykonania staje się zintegrowanie sejsmometrów z mediami społecznościowymi tak, aby sejsmometr mógł automatycznie wysyłać wiadomość, gdy zarejestrowałby wstrząs powyżej pewnej wartości. Po ilości wpisów (zgłoszeń) z danego obszaru można by natomiast wnioskować o sile trzęsienia ziemi i skali zniszczeń, oczywiście zakładając, że nie została zniszczona sieć komórkowa i po wystąpieniu trzęsienia użytkownicy mają nadal dostęp do internetu. Należy mieć na uwadze, że przywrócenie łączności jest obecnie jednym z priorytetów służb ratunkowych (Goldstein 2010). Pokaźna grupa badaczy opublikowała także prace dotyczące huraganów. Mandel i in. (2012), dzięki opracowanej przez siebie metodzie automatycznej klasyfikacji *tweedów*, badali korelację liczby wpisów z pomiarem siły huragan Irene. Na tej podstawie szacowali miejsca, które najbardziej potrzebują pomocy służb w danej chwili. Również Meier i in. (2013) skupili się w swoich badaniach nad rozwinięciem metod automatycznej klasyfikacji *tweedów*, dotyczących danego zjawiska. Shelton i in. (2014) badali reakcję użytkowników Twittera na uderzenie huraganu Sandy. Wykorzystali oni innowacyjne podejście badawcze z użyciem opracowanej przez Jessopa (2008) metody TPSN (*territory*, *place*, *scale*, *network*) i zaadaptowanej na potrzeby badań geograficznych. Metoda ta zakłada analizowanie wymienionych wymiarów łącznie, przez co unika się wielu metodologicznych błędów popełnianych przy analizowaniu danych tylko w jednym wymiarze.

Caragea i in. (2014) w swoich doświadczeniach badali za pomocą analizy semantycznej *tweetów* nastroje ludzi dotkniętych kataklizmem w kolejnych fazach działania służb ratunkowych. Wskazują oni, że pierwszymi ratownikami są osoby z sąsiedztwa, również dotknięte kataklizmem, dopiero w dalszej fazie wkraczają profesjonalne służby. Autorzy podkreślają użyteczność takich badań w opracowywaniu przyszłych procedur działania na wypadek klęsk żywiołowych. Chae i in. (2014) stwierdzają, że sposób raportowania o zjawisku zależy w znacznej mierze od tego zjawiska i jego wymiaru czasoprzestrzennego. Przykładowo podczas huraganu Sandy, który trwał ponad tydzień, dużo osób pisało o nim na obszarach potencjalnie narażonych na jego

uderzenie, a po przejściu huraganu liczba tweetów bardzo zmalała ze względu na duże zniszczenia, jakie huragan wyrządził w infrastrukturze. W przypadku tornada, które występuje nagle i ma mały zasięg przestrzenny, przed jego nadejściem w zasadzie nikt nie wspomina o zagrożeniu, natomiast po jego przejściu ilość *tweetów* na jego temat znacznie wzrasta nawet na obszarach nim dotkniętych. Oprócz wykorzystania Twittera do monitoringu i oceny skutków zdarzeń ekstremalnych, w światowej literaturze pojawiają się opracowania dotyczące mobilności ludzi. Wymienić tu można opracowania Hawełki i in. (2013), dotyczące migracji pomiędzy poszczególnymi państwami w skali globalnej oraz Lenormanda i in. (2014), w której autorzy badali wykorzystanie danych z Twittera i telefonii komórkowej do badania dziennej ścieżki życia mieszkańców dużych miast na przykładzie Barcelony i Madrytu. Badaniem mobilności mieszkańców dużych miast i integracją danych z mediów społecznościowych zajmowali się także Sagl i in. (2012). Ciekawe badania prowadzili również Frias-Martinez i in. (2012), którzy użyli *tweetów* do scharakteryzowania krajobrazu miejskiego. Krótko podsumowując powyższe badania, można stwierdzić, że Twitter jako źródło informacji odegrał w nich istotną rolę i przeprowadzenie takich badań bez dostępu do danych Twittera byłoby bardzo trudne, kosztowne, a nawet niemożliwe. Dzięki zastosowaniu metody *data miningu* w pozyskiwaniu informacji z mediów społecznościowych możliwe staje się prowadzenie badań z pogranicza wielu dyscyplin, w których centrum jest człowiek i jego mniej lub bardziej subiektywne postrzeganie otaczającego go świata.

Działanie serwisu i pozyskiwanie danych z Twittera

Z założenia wszystkie *tweety* jakie opublikuje użytkownik na swojej stronie profilowej są publiczne i dostępne dla każdego. Może on także dodać innego użytkownika do „obserwowanych", dzięki czemu na bieżąco będzie otrzymywał *tweety* publikowane przez niego na swoim koncie – stanie się jego „obserwującym". Co należy podkreślić, relacja obserwowany-obserwujący nie jest dwustronna. To odróżnia Twittera od innych mediów społecznościowych, jak Facebook, gdzie tworzy się kręgi znajomych. Obserwujący może polubić *tweet* osoby obserwowanej (wtedy ona dostanie o tym informację, jest to wyraz aprobaty dla opublikowanej treści), ma też możliwość przekazania *tweeta* dalej, swoim obserwującym, publikując go na swoim profilu. W treści *tweeta* zawrzeć można także słowa specjalne. Wpisując w treści *tweeta* @nazwa_użytkownika, można się zwrócić bezpośrednio do danej osoby. Drugi rodzaj słów specjalnych to tzw. *#hashtagi*. Mają one szczególnie duże znaczenie przy semantycznej analizie treści *tweetów*, ponieważ dzięki nim użytkownicy sami dokonują klasyfikacji *tweetów*. Są to poniekąd słowa kluczowe, które tematycznie łączą różne *tweety*. Naciśnięcie takiego słowa wyświetli inne wiadomości zawierające ten sam *#hashtag*.

Według Crooksa i in. (2013) klasyczny *data mining* sprowadza się do trzech zasadniczych operacji: wyodrębnienia danych od dostawcy (serwisy społecznościowe) przez aplikację dostępową opartą na programowalnym interfejsie (API), przeanalizowanie tych informacji, zintegrowanie ich i zapisaniu do bazy, a następnie analizę w programie GIS, wyodrębniającą badane zjawisko.

Wyodrębnienie danych od dostawcy

Twitter udostępnia ogólnodostępne API, które umożliwia pisanie własnych programów łączących się bezpośrednio z bazą Twittera. Dane można pobierać przez architekturę *Representational State Transfer* (REST), za pomocą zapytań *Hypertext Transfer Protocol* (HTTP). Jako metodę autoryzacji dostępu użyto standardu autoryzującego OAuth. Aby utworzona aplikacja mogła połączyć się z Twitterem, musi przejść procedurę rejestracji i uzyskać odrębny dla niej kod OAuth. Twitter udostępnia trzy zasadnicze rodzaje API: *Streaming API*, *REST API* i *Search API*. Najbardziej użyteczny z punktu widzenia *data miningu* wydaje się *Streaming API*, które zapewnia dostęp do publikowanych *tweetów* prawie w czasie rzeczywistym. Ponieważ przez *Streaming API* nie ma możliwości pobierania danych historycznych, konieczne jest, aby aplikacja je wykorzystująca pracowała na serwerze non stop. *Streaming API* komunikuje się za pomocą protokołu HTTP i wykorzystuje trzy główne komendy GET, POST i DELETE. Jako że dostęp przez *Streaming API* jest bezpłatny, ma pewne ograniczenia. Można pobrać jedynie 1% wszystkich tweetów opublikowanych na Twitterze w danej chwili. Wydawać się może, że 1% danych jest niewystarczający, ale należy mieć na uwadze, że *tweety* z geolokalizacją stanowią obecnie właśnie około 1% wszystkich publikowanych *tweetów* (Morstatter i in. 2013), a dodatkowo możemy także zawęzić obszar analizy, na przykład do kontynentu i w ten sposób mieć gwarancję, że większość *tweetów* z geolokalizacją zostanie wyłapana przez aplikację. *REST API* i *Search API* umożliwiają dostęp do danych historycznych, ich możliwości są jednak znacznie bardziej ograniczone. Oprócz metod bezpłatnych Twitter oferuje także płatną usługę dostępu do danych Firehose, która gwarantuje pozyskanie 100% danych. Morstatter i in. (2013) porównali *tweety* pozyskane zarówno przez Twitter API, jak również przez Firehose, i uzyskali zgodność na poziomie 90%, co potwierdza wysoką skuteczność *Streaming API*.

Integracja danych w bazie

Duże ilości danych zapisuje się i przechowuje w hurtowniach danych (*warehouse*). W tym celu wykorzystuje się relacyjne bazy danych. Dzięki relacjom można ograni-

czyć redundancję danych i zastosować słowniki oraz indeksy, co znacznie przyspiesza ich przeszukiwanie. Dodatkowo, na potrzeby analiz przestrzennych, najbardziej popularne bazy danych oferują możliwość zastosowania geometrycznego typu danych (*geometry type*), umożliwiając wykonywanie analiz przestrzennych bezpośrednio po stronie bazy danych.

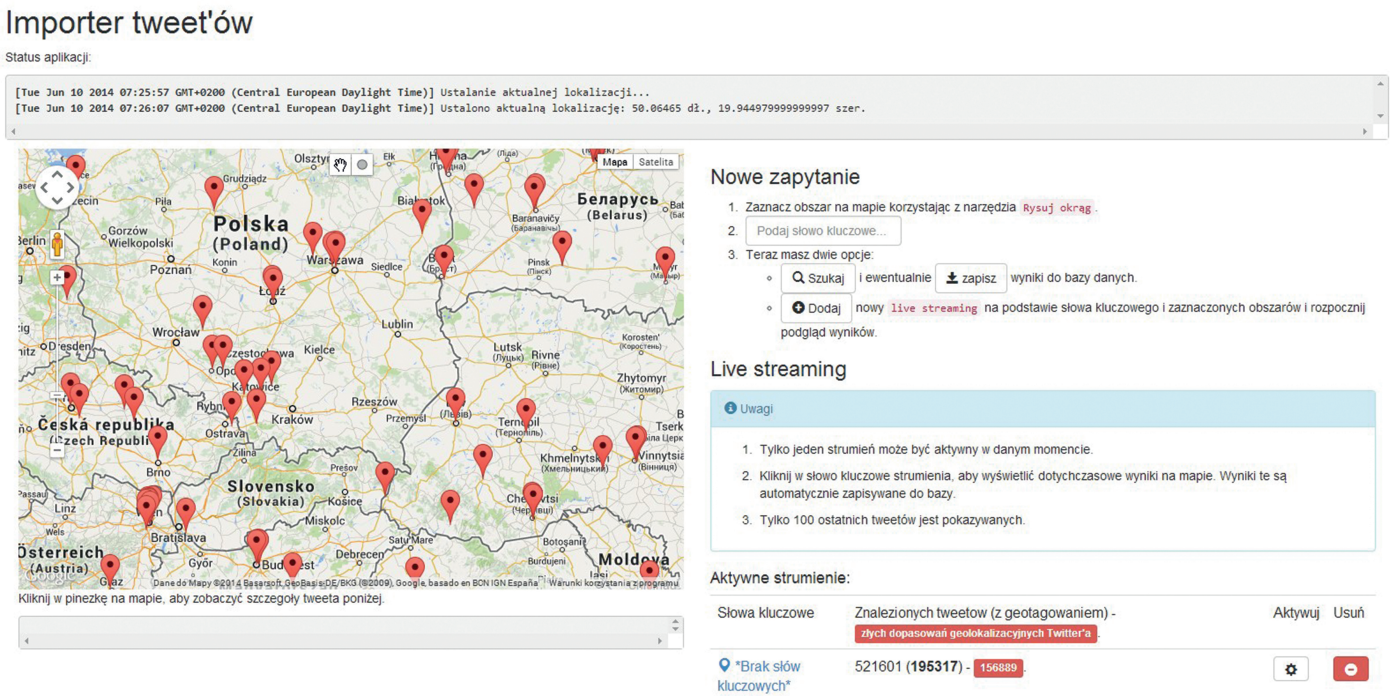
Właściwe analizy w oprogramowaniu GIS

Oprogramowanie GIS może być użyte zarówno do wykonywania analiz, jak i do wizualizacji ich wyniku. Jest to zależne od metody badacza – czy analizy przestrzenne wykonuje po stronie bazy danych i wtedy używa narzędzi GIS jedynie do opracowania danych kartograficznych przedstawiających rezultat badań, czy też wykonuje wszystkie analizy przestrzenne po stronie oprogramowania. Możliwa też jest metoda, w której po stronie bazy danych zostaną wykonane proste analizy, wymagające jednak przetworzenia dużej liczby danych (w tym celu baza danych ma znacznie większą wydajność), a w oprogramowaniu GIS bardziej zaawansowane analizy, na przykład wykorzystujące jeszcze inne źródła danych. Trzeba jednak zaznaczyć, że nie można postawić wyraźnej granicy między oprogramowaniem GIS rozumianym jako aplikacja desktop a bazą danych, ponieważ narzędzia te się przenikają i część narzędzi oprogramowania GIS można zintegrować z bazą danych.

Badania

Na potrzeby badań autor wraz z programistą Bartoszem Krupą opracował program komputerowy Importer Tweetów, który wykorzystuje *Streaming API* i pobiera *tweety* dla zdefiniowanego obszaru. Program umożliwia na wstępie filtrację danych i ograniczenie się do wybranych słów kluczowych, ale w chwili obecnej zbierane są wszystkie *tweety*, jakie udostępnia tą metodą Twitter, a selekcja odbywa się już po stronie bazy danych. Taka metoda jest lepsza przy analizie *big data*, gdyż lepiej pozyskać więcej danych i użyć do ich ekstrakcji metod *data miningu*, niż na wstępie zawężać strumień danych (Zafarani i in. 2014).

Aplikacja Importer Tweetów jest aplikacją działającą w przeglądarce internetowej, która została napisana w technologii *Java Script* na platformę Node.js. Interfejs graficzny aplikacji (ryc. 1) umożliwia zdefiniowanie obszaru analizy, słów kluczowych, a następnie aktywowanie strumienia *tweetów* i zapisywanie ich w czasie rzeczywistym do bazy danych. *Tweety* są w postaci formatu *JavaScript Object Notation* (JSON), który następnie jest przekształcany na typ danych *string* i zapisywany w bazie. Jako bazę danych wybrano darmową bazę PostgreSQL. Każdy *tweet* zapisywany w bazie



Ryc. 1. Interfejs graficzny aplikacji Importer Tweetów
Fig. 1. Graphic User Interface (GUI) of application Tweets Importer

Źródło: opracowanie własne.
Source: author's own study.

posiada 35 atrybutów, z których najważniejsze to oczywiście treść samego *tweeta*, dokładny czas jego publikacji, geolokalizacja w postaci współrzędnych geograficznych, pseudonimu autora. Oprócz tych atrybutów zapisywana jest również: informacja o języku *tweeta*, identyfikator *tweeta* źródłowego (w przypadku kiedy tweet jest odpowiedzią na innego *tweeta*), kraj z jakiego *tweet* został wysłany, rodzaj oprogramowania (np. Twitter dla Androida), a także informacje dotyczące konta autora (np. kiedy zostało założone, ile osoba ma obserwujących, a ile osób ją obserwuje). Należy dodać, że oprócz lokalizacji za pomocą współrzędnych geograficznych, każdy t*weet* posiada mniej dokładną lokalizację ograniczoną przeważnie do obszaru miasta bądź regionu, uzyskiwaną na podstawie IP komputera.

Tweety zapisywane są do bazy nierelacyjnej, gdzie jedna krotka to jeden *tweet*. Na potrzeby dalszych analiz baza jest konwertowana do bazy relacyjnej, usuwana jest redundancja danych, tworzone są klucze główne i obce, a przede wszystkim zakładane są indeksy i tworzona dodatkowa kolumna z geometrią. Dodatkowo, na potrzeby konkretnej analizy, używanie wszystkich 35 atrybutów nie jest konieczne, a więc dane są ograniczane. Do analiz stosuje się także ograniczenia czasowe i przestrzenne, dzięki czemu pomimo bardzo dużej ilości danych źródłowych używane są tylko te dane, które rzeczywiście powinny być analizowane, a pozostałe, które z całą pewnością można wykluczyć, nie są uwzględniane w analizach. Takie postępowanie znacznie poprawia wydajność generowania zapytań do bazy danych.

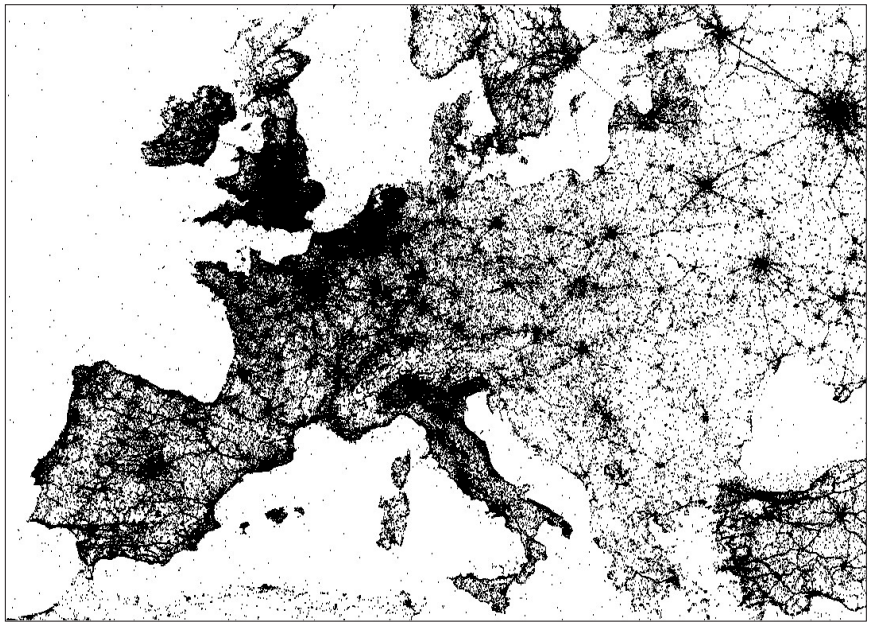
Domyślnie baza PostgreSQL nie wspiera danych przestrzennych, jednak dzięki dodatkowym rozszerzeniom wsparcie to jest możliwe. Do celów porównawczych i testowania wydajności zastosowano dwa równorzędne rozszerzenia: ArcSDE od firmy ESRI oraz darmowe rozszerzenie PostGIS. Dzięki tym rozszerzeniom możliwe staje się wykonywanie skomplikowanych analiz przestrzennych w języku *Structured Query Language* (SQL) z wykorzystaniem zarówno danych z Twittera, jak i danych pochodzących z innych źródeł, na przykład plików wektorowych z jednostkami podziału administracyjnego.

Do analiz wykorzystywane jest oprogramowanie ArcGIS for Desktop w wersji 10.2, które jest połączone z bazą danych PostgreSQL. Schemat pracy z danymi jest następujący: najpierw następuje przygotowanie danych pod konkretną analizę, odbywa się to za pomocą zapytań SQL wykonywanych w oprogramowaniu PgAdmin, stosowane są filtry przestrzenne, czasowe oraz ilościowe, następnie dane są filtrowane za pomocą atrybutów. Tak przygotowane dane są wyświetlane w oprogramowaniu ArcGIS, gdzie wykorzystywane są już typowe narzędzia GIS, jak na przykład łączenie chronologiczne *tweetów* danej osoby, aby prześledzić jej ścieżkę życia. Ostatnim etapem pracy z danymi jest ich wizualizacja graficzna.

Wstępne wyniki

Analizy można podzielić na dwa główne rodzaje: analizy na podstawie lokalizacji, badające skąd dana osoba pisze i jak się zmienia lokalizacja kolejnych *tweetów* w czasie, ale bez analizy treści tweetów, oraz analizy treści *tweetów* oparte na słowach kluczowych, lub analizie semantycznej. Ten drugi rodzaj analiz jest znacznie trudniejszy ze względu na zróżnicowanie językowe, synonimy i mnogość odmian przez przypadki słów kluczowych.

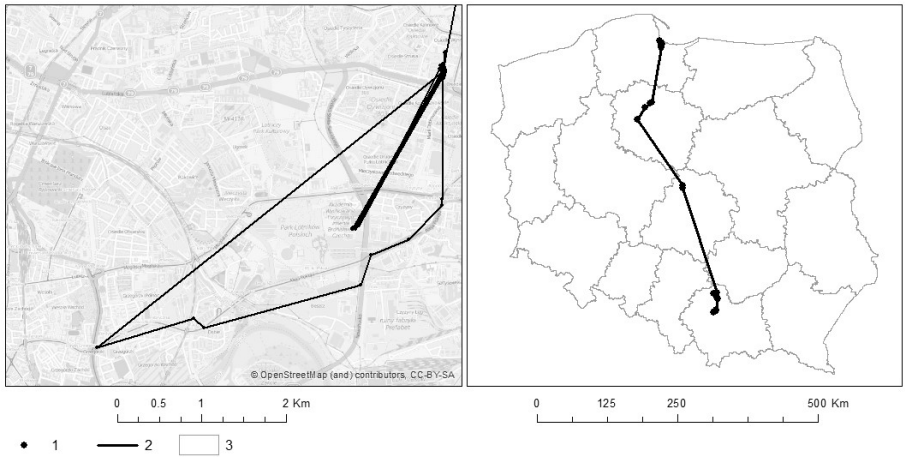
Na potrzeby niniejszej analizy (stan na luty 2015 roku) autor zebrał w bazie ponad 1 mld *tweetów* z obszaru Europy, Azji, Ameryki Północnej i Południowej. Dane są zbierane od 19 grudnia 2013 roku. Po zgeolokalizowaniu *tweetów* na mapie Europy (ryc. 2) można zauważyć znaczne dysproporcje pomiędzy poszczególnymi krajami w ich ilości, co spowodowane jest głównie różną popularnością tego serwisu w tych krajach. W Europie najbardziej aktywnymi krajami są Wielka Brytania, kraje Bene-



Ryc. 2. Tweety z geolokalizacją – pierwsze półrocze 2014 roku
Fig. 2. Tweets with geolocation – the first half of 2014 year
Źródło: opracowanie własne.
Source: author's own study.

luksu, Irlandia, aktywne są też północne Włochy i północna Francja. Stosunkowo dużą popularnością cieszy się Twitter w Turcji. Na rycinie 2 wyraźnie także widać szlaki komunikacyjne między dużymi miastami, co można tłumaczyć zwiększonymi potokami ludzi pokonujących te trasy w porównaniu z obszarami obok. Wyraźnie też zaznacza się linia wybrzeża. Obszary nadmorskie charakteryzują się dużą gęstością zaludnienia, dodatkowo są to często obszary turystyczne, co sprzyja zwiększonej liczbie wpisów na serwisach społecznościowych.

Wstępne analizy potwierdzają duże możliwości zaproponowanej metody badań. Przykładowo na rycinie 3 przedstawiono miesięczną ścieżkę życia losowo wybranej osoby. Widzimy, że osoba ta najprawdopodobniej mieszka w Krakowie, ponieważ tam publikuje najwięcej wpisów. Można zauważyć dwa obszary, gdzie występuje największa ich koncentracja, jedno z tych miejsc będzie zapewne miejscem zamieszkania, drugie to szkoła bądź miejsce pracy. Widzimy też, że osoba ta odbyła jednorazową podróż do Gdańska, a podróż ta odbywała się koleją, co potwierdza treść *tweetów*. Zestawiając podobne analizy dla większej liczby użytkowników i odpowiednio je wizualizując, możliwe jest zbadanie potoków pasażerskich w obrębie dużego miasta czy migracji pomiędzy metropoliami.



Ryc. 3. Przykładowa ścieżka życia mieszkańca Krakowa na podstawie danych zbieranych przez jeden miesiąc

Fig. 3. Sample of live path of Cracow's resident based on data collected for one month

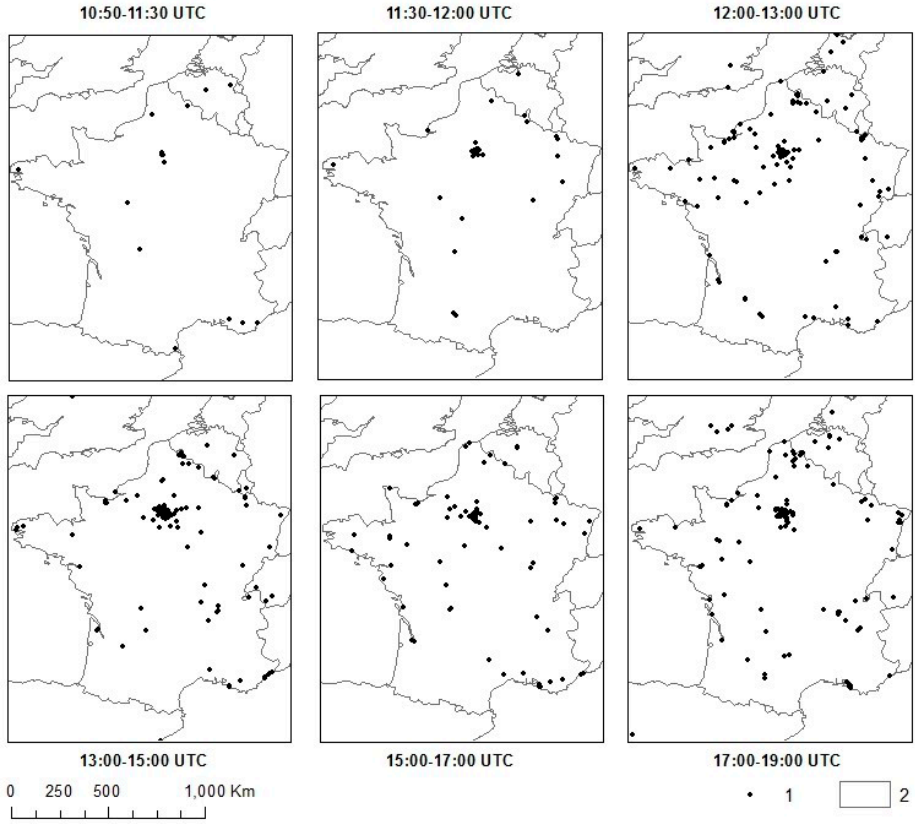
Objaśnienia: 1 – tweet, 2 – połączenia pomiędzy tweetami, 3 – granice województw

Explanations: 1 – tweet, 2 – connections between tweets, 3 – bondaries of provinces

Źródło: opracowanie własne.

Source: author's own study.

Analizy po słowach kluczowych treści *tweetów* pozwalają na wizualizację przestrzenną, określającą, co użytkownicy sądzą na określony temat. Przez użycie *#hashtagów* użytkownicy Twittera sami dokonują klasyfikacji treści, co znacznie ułatwia analizy. Na rycinie 4 przedstawiono wizualizację *tweetów* dla *hashtagu #charliehebdo*



Ryc. 4. Wizualizacja tweetów z hashtagiem #charliehebdo w kolejnych godzinach po zamachu w Paryżu 7 stycznia 2015 roku

Fig. 4. Visualization of tweets with hashtag #charliehebdo in the next hours after assassination in Paris – 7th of January 2015 year

Objaśnienia: 1 – tweet, 2 – granice państw

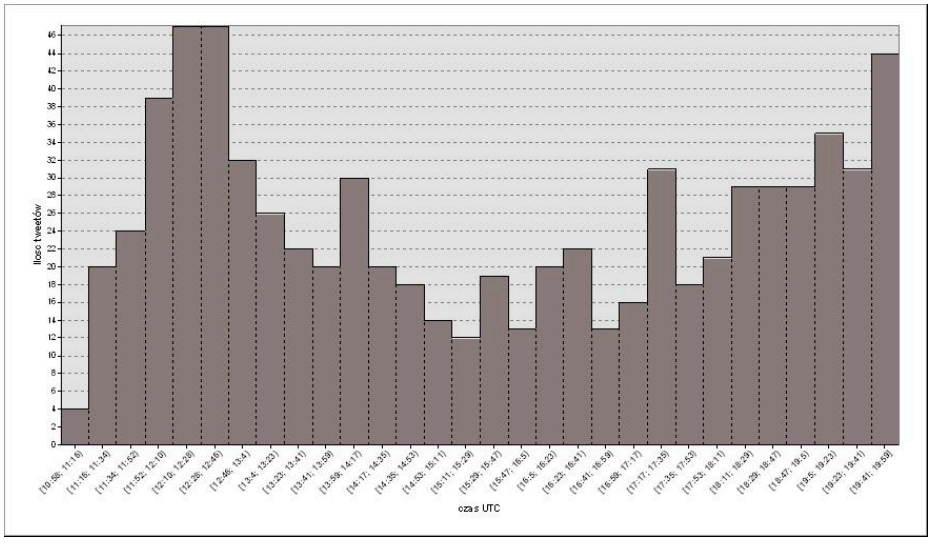
Explanations: 1 – tweet, 2 – borders

Źródło: opracowanie własne.

Source: author's own study.

w dniu 7 stycznia 2015 roku, po zamachu na redakcję czasopisma „Charlie Hebdo". Zamach nastąpił około godziny 10:30 czasu UTC, pierwszy *tweet* z geolokalizacją pojawił się o godzinie 10:58 z miejsca oddalonego 2,5 km od redakcji. Co ciekawe, z miejsca zamachu opublikowano tylko dwa *tweety* z geolokalizacją, co mogło być spowodowane blokadą tego miejsca przez jednostki policji i niedopuszczaniem osób postronnych w pobliże miejsca zamachu. Na rycinie 5 przestawiającej ilość tweetów z *hashtagiem #charliehebdo* w poszczególnych przedziałach czasowych widać wyraźnie dwa momenty kulminacyjne, kiedy informacji było najwięcej. Pierwszy trwa około dwie godziny po zamachu i są to głównie *tweety* opublikowane w rejonie Paryża, drugi natomiast przypada na godziny wieczorne, tradycyjną porę dzienników telewizyjnych i są to już *tweety* z całej Francji i innych krajów, z tym że wpisy z innych krajów stanowią bardzo niewielki odsetek wszystkich wpisów.

Oprócz analizy treści *tweetów* i ich lokalizacji zwizualizować można każdy inny atrybut *tweeta*. Bardzo ciekawe analizy dotyczące mniejszości zamieszkujących dany obszar oraz ruchu turystycznego można uzyskać, wizualizując język, w jakim *tweet*



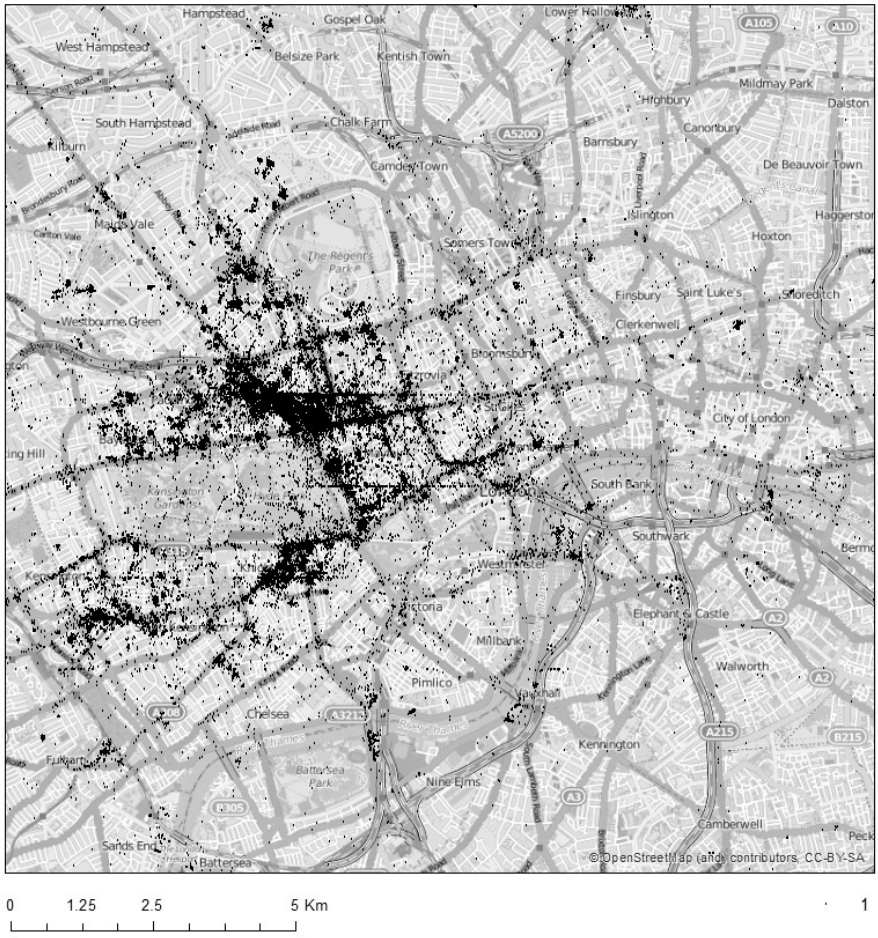
Ryc. 5. Ilość tweetów z hashtagiem #charliehebdo w kolejnych godzinach po zamachu w Paryżu 7 stycznia 2015 roku

Fig. 5. The amount of tweets with hashtag #charliehebdo in the next hours after assassination in Paris – 7th of January 2015 year

Źródło: opracowanie własne.

Source: author's own study.

został napisany. Na rycinie 6 zwizualizowano *tweety* w języku arabskim w Londynie. Widać znaczną ich koncentrację w rejonie ulicy Edgware Road, czyli miejscu zamieszkania społeczności arabskiej.



Ryc. 6. Wizualizacja tweetów w języku arabskim w Londynie

Fig. 6. Visualization of arabic language tweets in London

Objaśnienia: 1 – tweet

Explanations: 1 – tweet

Źródło: opracowanie własne.

Source: author's own study.

Dyskusja

Użycie mediów społecznościowych jako źródła informacji geograficznej otwiera nowe horyzonty w badaniach na pograniczu geografii i socjologii. Dzięki wykorzystaniu danych typu *big data* dostajemy ogromną ilość informacji, w analizie których ograniczeniem jest tylko pomysłowość badacza i możliwości techniczne komputerów.

Przykłady wymienione we wstępie potwierdzają dużą przydatność metod *data miningu*, ale do analizy *tweetów* nie mamy jeszcze narzędzia idealnego, co podkreślili Shelton i in. (2014). Potwierdzili oni, że gęstość wpisów dotyczących huraganu rośnie na obszarach najbardziej nim dotkniętym, jednocześnie wskazali, że należy do tej metody podchodzić ostrożnie, ponieważ może ona wskazywać nie tyle na zniszczenia, ile na nierówny dostęp do infrastruktury. Na przykład obszary najbardziej zniszczone, najbardziej potrzebujące pomocy będą pozbawione elektryczności, dlatego nie będzie z tego obszaru *tweetów*, co spowoduje powstanie błędnego wniosku, że obszary te tej pomocy właśnie nie potrzebują. W rezultacie metoda ta może tylko pogłębić nierówności społeczne. Autorzy podkreślają, że oprócz analizy ilościowej bardzo ważna jest analiza jakościowa treści *tweetów*, jednocześnie wskazują, że wystarczy niewielka próbka danych z wszystkich zebranych *tweetów*, aby wyciągnąć istotne statystycznie wnioski, ilość *tweetów* i tak będzie bowiem bardzo duża. Połączenie analizy jakościowej (polegającej na analizie treści przez badacza) z ilościową (wykorzystującą metodę *data miningu*) sprawia, że możliwe staje się lepsze zrozumienie treści *tweetów*, jednocześnie, dzięki analizie ilościowej, liczba *tweetów*, jakie trzeba przeanalizować jakościowo, maleje. Sama analiza ilościowa jest zbyt dużym uproszczeniem i nie bada zawiłości w relacjach społeczno-przestrzennych, a przeanalizowanie danych jakościowo, bez metod ilościowych jest, ze względu na ogrom pracy, niewykonalne. Autorzy wskazują na potrzebę wykorzystania już istniejących metod badających relacje społeczno-przestrzenne, jak chociażby metoda TPSN zaproponowana przez Jessopa i in. (2008), która zależność pomiędzy ilością *tweetów* na danym obszarze a skalą badanego zjawiska tłumaczy bardziej holistycznie niż prosta wizualizacja w przestrzeni kartezjańskiej *tweet-to-point*. Ich zdaniem ważniejsze od analizy w czasie rzeczywistym jest wykonanie analiz po wystąpieniu zjawiska, aby dzięki temu lepiej to zjawisko zrozumieć, poznać reakcje ludzi w jego obliczu i na tej podstawie wypracować wzorce postępowania, aby w przyszłości przy wystąpieniu podobnych zjawisk ograniczyć jego skutki i rozmiar strat.

Podsumowanie

Wykorzystanie mediów społecznościowych, w szczególności Twittera, oraz nowoczesnych metod analizy danych otwiera szereg nowych możliwości w badaniach społeczno-geograficznych, pozwalając rzucić nowe spojrzenie na stare problemy badawcze bądź też zbadać problemy, które dotychczas były niemożliwe do zbadania z powodu braku, albo niedostatecznej liczby, wiarygodnych danych. Otwarte API Twittera sprawia, że dostęp do danych ma w zasadzie każdy bez potrzeby płacenia za nie, znika więc podstawowa bariera przy zbieraniu danych społecznych – duży koszt badań ankietowych.

Liczne światowe badania potwierdzają, że Twitter stanowi cenne źródło wiedzy w przypadku zjawisk ekstremalnych, a także pozwala lepiej zrozumieć te zjawiska i utworzyć procedury, które lepiej będą chronić przed nimi ludzi. Należy jednak zauważyć, że nie jest to źródło danych pozbawione wad. Najlepiej sprawdzi się integracja tradycyjnych metod monitoringu z rozwiązaniami opartymi o *data mining*. Badania nad wykorzystaniem Twittera jako źródła nieświadomej informacji geograficznej są w toku, i należy się spodziewać, że w niedalekiej przyszłości algorytmy oparte na sztucznej inteligencji będą już na tyle dobre, że ręczna klasyfikacja *tweetów* przez człowieka nie będzie wymagana, co zwiększy wydajność tych metod i przełoży się na ich skuteczność.

Badania zostały sfinansowane w ramach działalności statutowej Instytutu Geografii i Gospodarki Przestrzennej Uniwersytetu Jagiellońskiego w latach 2013–2014.

Literatura

Alexander D., 2014, *Social Media in Disaster Risk Reduction and Crisis Management*, Science and engineering ethic, 20 (3), 717–733.

Anderson C., 2008, *The end of theory: The data deluge makes the scientific method obsolete*, Wired, Mag. 15 (7).

Caragea C., Squicciarini A., Stehle S., Neppalli K., Tapia A., 2014, *Mapping Moods: Geo-Mapped Sentiment Analysis During Hurricane Sandy*, [w:] S.R. Hiltz, M.S. Pfaff, L. Plotnick, A.C. Robinson (red.), *Proceedings of the 11th International ISCRAM Conference – University Park*, Pennsylvania, http://www.wired.com/2008/06/pb-theory/ (30.06.2015).

Chae J., Thom D., Jang Y., Kim S., Ertl T., Ebert D., 2014, *Public behavior response analysis in disaster events utilizing visual analytics of microblog data*, Computer & Graphic, 38, 51–60.

Cheong F., Cheong C., 2011, *Social media data mining: A social network analysis of tweets during the 2010–2011 Australian floods*, PACIS 2011 Proceedings, 6–46.

Christensen C., 2011, *Twitter revolutions? Addressing social media and dissent*, The Communication Review, 14, 155–57.

Crampton J., Graham M., Poorhuis A., Shelton T., Stephens M., Wilson M., 2013, *Beyound the geotag: Situating 'big data' and leveraging the potential of the geoweb*, Cartography and Geographic Information Science, 40 (2), 130–139.

Crooks A., Croitoru A., Stefanidis A., Radzikowski J., 2013, *#Earthquake: Twitter as a Distributed Sensor System*, Transactions in GIS, 17 (1), 124–147.

de Longueville B., Annoni A., Schade S., Ostlaender N., Whitmore C., 2010, *Digital Earth's Nervous System for crisis events: Real-time Sensor Web Enablement of Volunteered Geographic Information*, International Journal of Digital Earth, 3 (3), 242–259.

Dorsey J., 2006, *Just setting up my twttr*, https://twitter.com/jack/status/20 (15.01.2015).

Frias-Martinez V., Soto V., Hohwald H., Frias-Martinez E., 2012, *Characterizing Urban Landscapes Rusing Geolocated Tweets*, [w:] *Proceedings, 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, 239–248.

Goldstein H., 2010, *Engineers Race to Restore Communications after Haiti Quake*, IEEE Spectrum, http://spectrum.ieee.org/tech-talk/telecom/internet/engineers-race-to-restore-communications-after-haiti-quake (20.10.2014).

Goodchild M.F., 2007, *Citizens as sensors: The world of volunteered geography*, GeoJournal, 69 (4), 211–221.

Hawełka B., Sitko I., Beinat E., Sobolevsky S., Kazakopoulos P., 2013, *Geo-located twitter as a proxy for global mobility patterns*, Cartography and Geographic Information Science, 41 (3), 260–271.

Hey T., Tansley S., Tolle K., 2009, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmonton, Washington, http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf (13.11.2014).

Hyvärinen O., Saltikoff E., 2010, *Social Media as a Source of Meteorological Observations*, Monthly Weather Review, 138 (8), 3175–3184.

Jessop B., Brenner N., Jones M., 2008, *Theorizing sociospatial relations*, Environment and Planning D: Society and Space, 26 (3), 389–401.

Kaplan A., 2012, *If you love something, let it go mobile: Mobile marketing and mobile social media 4x4*, Business Horizons, 55, 129–139.

Kitchin R., 2013, *Big data and human geography: Opportunities, challenges and risks*, Dialogues in Human Geography, 3 (3), 262–267.

Lenormand M., Picornell M., Garcia Cant'u O., Tugores A., Louail T., 2014, *Cross-checking different source of mobility information*, PLoS One 2014, 9 (8), http://dx.doi.org/10.1371%2Fjournal.pone.0105184 (16.12.2014).

Mandel, B., Culotta, A., Boulahanis, J., Stark, D., Lewis, B., Rodrigue, J., 2012, *A demographic analysis of online sentiment during hurricane Irene*, [w:] *LSM, 12 Proceedings of the Second Workshop on Language in Social Media*.

Megapanel PBI/Genius, 2014, http://www.panel.pbi.org.pl/megapanel.php (3.06.2014).

Meier P., Castillo C., Imran M., Elbassuoni S.M., Diaz F., 2013, *Extracting information nuggets from disaster-related messages in social media*, [w:] T. Comes, F. Fiedrich, S. Fortier, J. Geldermann, L. Yang (red.), *Proceedings of the 10th International ISCRAM Conference – Baden-Baden, Germany, May 2013*, 1–10.

Morstatter F., Pfeffer J., Liu H., Kathleen M., Carley M., 2013, *Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose*, Proceedings of ICWSM, http://arxiv.org/abs/1306.5204 (5.02.2015).

Newsweek, 2009, *A Twitter Timeline of the Iran Election*, http://www.newsweek.com/twitter-timeline-iran-election-80867 (21.01.2015).

O'Reilly T., Battelle J., 2009, *Web Squared: Web 2.0 Five Years On*, O'Reilly Media, Inc.

Sagl G., Resch B., Hawelka B., Beinat E., 2012, *From Social Sensor Data to Collective Human Behaviour Patterns: Analysing and Visualising Spatio-Temporal Dynamics in Urban Environments*, [w:] *Proceedings of the GI-Forum 2012: Geovisualization, Society and Learning*, 54–63.

Schade S., Díaz L., Ostermann F., Spinsanti L., Luraschi G., Cox S., Nuñez M., 2011, *Citizen-based sensing of crisis events: Sensor web enablement for volunteered geographic information*, Applied Geomatics, 5 (1), 3–18.

Schade S., Ostermann F., Spinsanti L., Kuhn W., 2012, *Semantic Observation Integration*, Future Internet, 4, 807–829.

Shelton T., Poorthuis A., Graham M., Zook M., 2014, *Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'*, Geoforum, 52, 167–179.

Sui D., Goodchild M., 2011, *The convergence of GIS and social media: Challenges for GIScience*, International Journal of Geographical Information Science, 25 (11), 1737–1748.

Zafarani R., Ali Abbasi M., Liu H., 2014, *Social Media Mining – An Introduction*, Cambridge University Press, Cambridge.

Zook M., Graham M., Shelton T., Gorman S., 2010, *Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake*, World Medical & Health Policy, 2 (2), 7–33.

Łukasz Beluch
Instytut Geografii i Gospodarki Przestrzennej
Uniwersytet Jagielloński
ul. Gronostajowa 7, 30-387 Kraków
e-mail: lbeluch@gis.geo.uj.edu.pl